This appendix is organized as follows.

- In Section A, we show implementation details for LLaVA-1.5-HD (Sec. A.1), data and prompts (Sec. A.2), and hyperparameters (Sec. A.3).
- In Section B, we present more qualitative results for response format prompts (Sec. B.1), compositional capabilities (Sec. B.2).
- In Section C, we discuss limitations with more details.

A. Implementation Details

A.1. LLaVA-1.5-HD

A.1.1 Preprocessing

Overview. We use CLIP-ViT-L-14 (224^2) as the base image encoder. We first select and pad the input image to a target resolution that effectively captures its details, and split the image into 224^2 grids. All 224^2 image patches are encoded by the CLIP image encoder separately and their features are merged back to a single large feature map. We then post-process the resulting feature map to a flattened list of features. We additionally concatenate the features of a fixed-resolution image to provide the model with a global context.

Target resolution selection. We predefine a set of resolutions to support up to six grids (1x1, 1x2, 1x3, 1x4, 1x5, 1x6, 2x2, 2x3, and their transpose). This system allows for a maximum resolution of 672x448 (or 448x672). Two criteria are enforced in the target resolution selection: (1) *Detail preservation*: the selected resolution preserves as much detail from the original image as possible; (2) *Resource efficiency:* the resolution should not be excessively large to avoid unnecessary consumption of pixels and memory (*e.g.* it should not select 448^2 for a 224^2 input image).

Postprocessing. We perform three steps of postprocessing to ensure that the final features can be processed effectively and efficiently by the language model. (1) *Padding removal.* Features corresponding exclusively to the paddings are discarded. This reduces the number of visual tokens processed by the language model and improves the efficiency. (2) *Row-end Tokens.* We append a special token to the end of each row of features, to provide an explicit indication of the shape of the image. Unlike the original LLaVA and LLaVA-1.5 that uses a fixed resolution, we now use a variable resolution for the image features of LLaVA-1.5-HD, such indication allows the language model to capture the exact shape and the size of the image for each sample. (3) *Flattening.* Finally, we flatten the image feature map and feed it into the language model along with language token features.

Data	Size	Response formatting prompts
LLaVA [9]	158K	–
ShareGPT [14]	40K	–
VQAv2 [4]	83K	Answer the question using a single word or phrase.
GQA [5]	72K	
OKVQA [11]	9K	
OCRVQA [12]	80K	
A-OKVQA [13]	66K	Answer with the option's letter from the given choices directly.
TextCaps [15]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [6, 10]	48K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG [7]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

Table 1. **Instruction-following Data Mixture** of LLaVA-1.5.

Data	Response formatting prompts
LLaVA-Bench, MM-Vet	–
VQAv2, GQA, TextVQA, MME, POPE	Answer the question using a single word or phrase.
ScienceQA, MMBench, SEED-Bench	Answer with the option's letter from the given choices directly.
VizWiz	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.

Table 2. **Response format prompt** for evaluation.

A.1.2 Training

Since we compute the visual features on the original 224^2 resolution that the vision encoder is trained on, we do not perform additional pretraining. We also do not perform additional high resolution pretraining for the visual projectors, and perform visual instruction tuning directly on the higher-resolution images.

A.2. Data

Our final training data mixture contains a variety of datasets: VQA [4, 5, 11, 13], OCR [12, 15], region-level VQA [6, 7, 10], visual conversation [9] and language conversation [14] data. We adopt multiple strategies to reduce training cost and enhance efficiency, detailed as follows:

1. For all VQA datasets, QA pairs from the same training image are merged into a single conversation.
2. For ShareGPT [14], we filter out invalid conversations as [2]. Unlike Vicuna [2], long conversations that surpass 2048 tokens are truncated rather than splitting to multiple conversations. This results in ∼40K conversations.
3. Each QA pair in A-OKVQA [13] is augmented k times, where k is the number of choices per question, to counterbalance the lack of multiple-choice data.
4. 80K conversations are sampled from OCRVQA [12].

5. For Visual Genome, we sample 10 annotations for images with additional annotations.
6. For RefCOCO, conversations are dissected into segments, each containing fewer than 10 conversations.
7. We obverse that language conversations are often longer than visual ones. For each batch, we sample conversations only from a single modality, and this speeds up the training by 25%, and does not affect the final outcome.

All data splits are concatenated together and sampled with the same probability. We present the response formatting prompts of the final instruction-following data mixtures in Table 1 and the response format prompts used for each evaluation benchmark in Table 2.

A.3. Hyperparameters

The latest Vicuna v1.5 [16] is used as the base LLM. LLaVA-1.5 uses the same set of hyperparameters as the original LLaVA, except that we halve the learning rate in pretraining due to the usage of the MLP projection layer instead of the original linear projection layer design. We show the training hyperparameters for both first-stage vision-language alignment pretraining and the second-stage visual instruction tuning in Table 3. We use greedy decoding for evaluation to ensure reproducibility.

Hyperparameter	Pretrain	Finetune
batch size	256	128
lr	1e-3	2e-5
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
epoch	1	
optimizer	AdamW	
DeepSpeed stage	2	3

Table 3. **Hyperparameters** of LLaVA-1.5 are the same as the original LLaVA, except that we halve the learning rate in pretraining due to the usage of the MLP projection layer.

B. Qualitative Results

B.1. Response Format Prompts

We show additional examples of LLaVA-1.5 generalizing to different unseen response format prompts.

First, as shown in Table 4, LLaVA-1.5 can provide details at different granularities in response to user's requests. When requested by the user, it is also capable of switching between response formats within the conversations.

Second, we provide another example of the constrained prompting to generate the prompts for Stable Diffusion models. We show an example of generating anime prompts in Table 6.

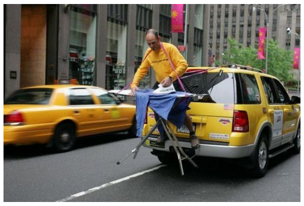
Visual input example, Different Format Prompts:	
	
User	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**
LLaVA-1.5	Yellow
User	Is this unusual? **Please explain briefly with a short sentence.**
LLaVA-1.5	Yes, it is unusual to see a man ironing clothes while standing on the back of a moving car.
User	Is this unusual? **Please explain in detail.**
LLaVA-1.5	Yes, this is an unusual scene. A man is ironing clothes while standing on the back of a moving yellow SUV. This is not a typical activity one would expect to see on a city street, as ironing clothes is usually done indoors, in a more private setting. The man's actions are not only unconventional but also potentially dangerous, as standing on the back of a moving vehicle can be hazardous and may lead to accidents or injuries. Additionally, the presence of a suitcase and a handbag in the scene suggests that the man might be traveling or on the go, which further emphasizes the unusual nature of the situation.
User	What is unusual about this image? **Answer briefly with a single sentence.**
LLaVA-1.5	A man is ironing clothes in the back of a car.
User	**Give a more detailed explanation.**
LLaVA-1.5	The image shows a man ironing clothes in the back of a car, which is an unusual sight as ironing is typically done indoors. The man is standing on a folding table, and the car appears to be a taxi. The scene takes place on a city street, with other cars and a truck visible in the background.

Table 4. LLaVA-1.5 learns to format the response according to the user's request, generalizes to unseen format instructions, and can alter the response format within the conversation upon the user's request.

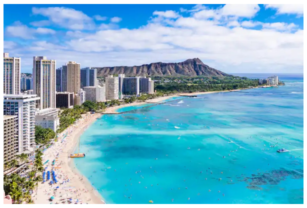
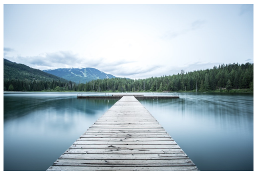
Visual input example, Visual Writing Task:	
	
User	Compose an engaging travel blog post about a recent trip to **this place**, highlighting **cultural experiences** and **must-see attractions**.
LLaVA	In my recent trip to Honolulu, I had the opportunity to experience the vibrant local culture and explore the beautiful beaches that this city has to offer. One of the must-see attractions was Waikiki Beach, where I witnessed a lively atmosphere with people sunbathing, swimming, and playing beach games. The crystal blue waters of the Pacific Ocean were a stunning sight, and the white sandy beach provided the perfect spot for relaxation. Aside from the beach, I also had the chance to immerse myself in the local culture by visiting the Pearl Harbor Memorial, which paid tribute to the events that led to the United States' involvement in World War II. The memorial was a somber reminder of the past, and it was fascinating to learn about the history surrounding the attack on Pearl Harbor. Other cultural experiences included trying local Hawaiian cuisine, such as poke bowls and traditional Hawaiian dishes, which were delicious and offered a unique taste of the region. I also had the opportunity to explore the city's bustling streets, where I came across local shops, markets, and street performers, adding to the lively atmosphere. Overall, my trip to Honolulu was an enriching experience, filled with cultural discoveries, breathtaking natural beauty, and unforgettable memories.
LLaVA-1.5	Title: A Sunny Escape to the Tropical Paradise of Waikiki Beach Introduction: I recently had the pleasure of visiting the beautiful Waikiki Beach in Hawaii, and I must say, it was an unforgettable experience. The crystal blue waters, the lively atmosphere, and the stunning cityscape in the background made it a perfect destination for a relaxing getaway. In this travel blog post, I'll share my experiences and must-see attractions during my trip to Waikiki Beach. Day 1: Arrival and Exploration Upon arrival, I was greeted by the warm and welcoming Hawaiian culture. The vibrant colors and lively music set the tone for my adventure. I decided to explore the area around Waikiki Beach, starting with the iconic Diamond Head Crater. The hike to the top was a bit challenging, but the breathtaking views of the beach and the city made it worthwhile. Day 2: Beach Day and Cultural Experiences I woke up early and headed to the beach to enjoy a day of sunbathing and swimming. The beach was bustling with activity, and I loved watching the surfers catching waves. In the evening, I attended a traditional Hawaiian luau, where I learned about the local customs and traditions, including hula dancing and the art of making leis. Day 3: City Tour and Sunset Cruise On the third day, I took a guided city tour to learn more about the history and culture of Hawaii. I visited the Pearl Harbor Memorial, the Bishop Museum, and the Iolani Palace, which was once the residence of the Hawaiian monarchs. In the evening, I went on a sunset cruise around the Waikiki coastline, where I enjoyed a delicious meal and witnessed the stunning sunset over the ocean. Day 4: Snorkeling and Relaxation For my last day in Waikiki, I decided to go snorkeling in the crystal clear waters of Hanauma Bay. The marine life was incredible, and I had the opportunity to see colorful fish and coral reefs up close. After a long day of exploring, I relaxed on the beach and soaked up the sun. Conclusion: My trip to Waikiki Beach was an unforgettable experience filled with adventure, culture, and relaxation. From the breathtaking views of the ocean to the vibrant cityscape, this tropical paradise has something for everyone. I highly recommend visiting Waikiki Beach for a sunny escape and a chance to immerse yourself in the beautiful Hawaiian culture.

Table 5. LLaVA-1.5 provides more detailed, visually-grounded responses for writing tasks with visual inputs than LLaVA.



Query: What are the things I should be cautious about when I visit here?

Reference (English): When visiting the pier over the lake, there are a few things you should be cautious about. First, ensure that you have appropriate footwear with good traction, as the pier may be wet or slippery, especially if it has been raining or if there is algae on the surface. Second, be mindful of the weather conditions, as the image shows a cloudy sky, which could indicate potential rain or storms. It is essential to stay safe and avoid being caught in a sudden storm or lightning. Lastly, be aware of your surroundings and any potential hazards, such as the presence of boats or other watercraft in the area. Always follow any posted signs or guidelines to ensure a safe and enjoyable experience.

Spanish: ¿Cuáles son las cosas de las que debería tener cuidado cuando visite aquí?

Cuando visite el lugar con la pier en el lago, debería tener cuidado de las siguientes cosas:

1. Mantenerse alejado del borde de la pier para evitar caer al agua o sufrir una lesión.
2. No subir a la pier si está prohibido o si no está diseñada para ello.
3. No arrojar objetos o desperdiciar en el agua, ya que esto puede dañar el ecosistema del lago.
4. No interferir con la fauna local, como los pájaros que se pueden encontrar en la zona.

Si decide nadar en el lago, asegúrese de conocer las condiciones del agua, como la profundidad y la presencia de algas, y de tener las habilidades y equipos adecuados para hacerlo de manera segura.

Tenir en cuenta estas precauciones ayudará a garantizar una experiencia segura y responsable en el lugar.

Translation: When visiting the place with the pier at the lake, you should be careful about the following things:

1. Stay away from the edge of the pier to avoid falling into the water or getting injured.
2. Do not climb onto the pier if it is prohibited or not designed for it.
3. Do not throw objects or litter into the water, as this can harm the lake's ecosystem.
4. Do not interfere with the local wildlife, such as birds that may be found in the area.

If you decide to swim in the lake, make sure to be aware of water conditions such as depth and the presence of algae, and have the appropriate skills and equipment to do so safely. Taking these precautions will help ensure a safe and responsible experience at the location.

Japanese: ここを訪れる際に注意すべきことは何ですか？

この場所は、寂しい湖に面した木製の桟橋です。訪れる際に注意すべきことは、桟橋の安全性と周囲の環境に対する配慮です。桟橋の構造が安定していることを確認し、桟橋が浸水していないことを確認することが重要です。また、湖の水深や周囲の地形に注意して、安全に桟橋を渡ることができるようにすることも大切です。さらに、湖の水質や周囲の生態系に配慮して、環境に負荷をかけないようにすることも重要です。

Translation: This place is a wooden pier facing a serene lake. When visiting, it's important to pay attention to the safety of the pier and respect the surrounding environment. Confirming the stability of the pier's structure and ensuring that it is not submerged in water are crucial. Additionally, being mindful of the lake's depth and the surrounding terrain is important to safely cross the pier. Furthermore, it's essential to consider the water quality of the lake and the local ecosystem to minimize any impact on the environment.

Korean: 여기를 방문할 때 주의해야 할 사항은 무엇인가요?

이 장소는 물체 위의 부두로, 맑고 푸른 물체에 있습니다. 방문자들은 이 부두에서 물체를 감상하고, 물체 주변의 자연 경치를 즐길 수 있습니다. 그러나 이 부두는 물체 위에 있기 때문에, 물체의 수준이 높아질 수 있으므로, 방문자들은 물체의 수준을 확인하고 적절한 예방 조치를 취해야 합니다. 또한, 부두에서 물체로 뛰어들지 않도록 주의해야 합니다. 이 부두는 매우 아름다운 자연 경치를 제공하므로, 방문자들은 이 장소를 즐기면서 안전한 방법으로 물체를 감상할 수 있도록 주의해야 합니다.

Translation: This place is located on a pier above clear and blue water. Visitors can admire the object from this pier and enjoy the natural scenery around the object. However, since the pier is above the object, the object's level can be high, so visitors should check the object's level and take appropriate precautions. Also, be careful not to jump from the pier into the object. This pier offers incredibly beautiful natural scenery, so visitors should enjoy this place while being mindful of safe ways to appreciate the object.

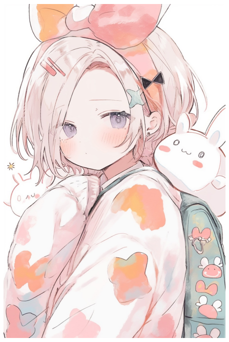
Figure 1. **Compositional capability: multilingual visual conversation.** LLaVA-1.5 generalizes to multilingual visual conversations, when training on visual instruction following data (English-only) together with the text-only ShareGPT data (multilingual). However, there can still be errors in some languages (*e.g.* Korean, errors marked in red).

B.2. Compositional Capabilities

We present qualitative examples of the compositional capabilities of LLaVA-1.5. As shown in Fig. 1, LLaVA-1.5 is capable of participating in multilingual visual conversations and adapting its output language based on the user's input, even though it has not been trained on multilingual visual instruction data. We hypothesize this emerging bahavior is a compositional capability learned from visual conversations (English-only) and the text-only ShareGPT data (multilingual). However, there can still be errors in some languages (*e.g*. Korean), which could be improved by incorporating more of those language data.

Additionally, in Table 5, we show another observed compositional capability after including the ShareGPT data in training. LLaVA-1.5 is able to produce more detailed and visually-grounded responses in writing tasks with visual inputs than LLaVA.

Visual input example, Constrained SD Prompt Generation:



User	Here's a formula for a Stable Diffusion image prompt: a cartoon-style image of [adjective] [subject] [haircut] [clothing] [doing action], [color scheme], detailed, trending on artstation. Generate the prompt for this image.
LLaVA-1.5	A cartoon-style image of a young girl with a bob haircut, wearing a white kimono, holding a stuffed animal, and posing for a picture, with a color scheme of pink and white, detailed, and trending on artstation.

Table 6. Constrained prompt generation for Stable Diffusion. Corresponding components are marked in color.

C. Limitations

Despite the promising results demonstrated by LLaVA-1.5, several limitations must be acknowledged. First, LLaVA-1.5 utilizes full image patches, potentially prolonging each training iteration. While visual resamplers [1, 3, 8] reduce the number of visual patches in LLMs, they currently cannot achieve convergence as efficiently as LLaVA with a comparable amount of training data, probably due to more trainable parameters in the resamplers. The development of a sample-efficient visual resampler could pave the way for future scaling-up of instruction-following multimodal models. Second, LLaVA-1.5 is not yet capable of processing multiple images due to the lack of such instruction-following data, and the limit of the context length. Third, although LLaVA-1.5 exhibits proficiency in following complex instructions, its problem-solving capabilities can still be limited in certain domains, which could be improved with a more capable language model and with high-quality, targeted visual instruction tuning data. Finally, despite its significantly reduced propensity for hallucination, LLaVA-1.5 is not exempt from producing hallucinations and occasionally disseminating misinformation, and should be used with caution in critical applications (*e.g*. medical).

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 5

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 5

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 1

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[10] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[11] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[12] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 1

[13] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 1

[14] ShareGPT. `https://sharegpt.com/`, 2023. 1

[15] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 1

[16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2